

Tomasz Limisiewicz

NLP & LLM Researcher

Researcher working on large language models, with a focus on tokenization, byte-level modeling, multilinguality, and multimodality. My work spans understanding model representations and developing methods for improving language models across languages and modalities.



tomlimi.github.io



tomlimi



LinkedIn



t.limisiewicz@gmail.com

EXPERIENCE

META | POSTDOCTORAL RESEARCHER

UNIVERSITY OF WASHINGTON | VISITING RESEARCHER

MAR 2025 -- PRESENT • Seattle, USA

Research with Luke Zettlemoyer's group on tokenization, byte-level language modeling, multilinguality, and multimodality. My recent work generalized the compute-optimal scaling laws for language model pre-training across different tokenizers (Compute Optimal Tokenization).

CHARLES UNIVERSITY | DOCTORAL RESEARCHER

MAR 2020 -- DEC 2024 • Prague, Czechia

Research on interpretability and control of language models, multilingual transfer, and bias mitigation.

INFERMEDICA | NLP / DATA SCIENTIST

NOV 2018 -- MAR 2020 • Wrocław, Poland

Developed a chatbot for medical interviewing and clinical inference.

SAMSUNG R&D | NLP ENGINEER

OCT 2017 -- SEP 2018 • Warsaw, Poland

Developed recursive neural models for handwriting recognition and machine translation.

AWARDS

Czech Republic Grant Agency Research Grant

2023

NAACL Outstanding Paper Award

2022

International Linguistics Olympiad Finalist

2011

SELECTED PUBLICATIONS

Compute Optimal Tokenization

preprint 2026

Tomasz Limisiewicz, Artidoro Pagnoni, Srini Iyer, Mike Lewis, Sachin Mehta, Alisa Liu, Margaret Li, Gargi Ghosh, Luke Zettlemoyer

MYTE: Morphology-Driven Byte Encoding for Better and Fairer Multilingual Language Modeling

ACL 2024

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, Luke Zettlemoyer

Debiasing Algorithm through Model Adaptation

ICLR 2024

Tomasz Limisiewicz, David Mareček, Tomáš Musil

A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank

NAACL 2022

Dan Malkin, Tomasz Limisiewicz, Gabriel Stanovsky
Outstanding Paper Award

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

2022

BigScience Workshop (Model Evaluation Initiative)

Introducing Orthogonal Constraint in Structural Probes

ACL 2021

Tomasz Limisiewicz, David Mareček

Full list of publications on Google Scholar (3,000+ citations)

EDUCATION

CHARLES UNIVERSITY

PH.D. COMPUTATIONAL LINGUISTICS

2020 -- 2024

Prague, Czechia

Thesis: Interpreting and Control in Linguistic Features in Multilingual Language Models

UNIVERSITY OF WROCŁAW

B.Sc. & M.Sc. MATHEMATICS

2012 -- 2017

Wrocław, Poland

WROCŁAW UNIVERSITY OF SCIENCE AND TECHNOLOGY

B.ENG. ELECTRONIC ENGINEERING

2011 -- 2016

Wrocław, Poland

SKILLS

Research

LLMs • Tokenization • Byte-level Models • Multilingual NLP • Interpretability

Languages

Polish (native) • English (fluent)
German, Czech (communicative)

SERVICE

Area Chair

ACL

Reviewer

NeurIPS • ICLR • ACL • COLM

Main Organizer

Tokenization Workshop
at ICML 2025 • COLM 2026